REMARKS

This amendment responds to the Final Office Action mailed August 20, 2007. In the Final Office Action the Examiner:

- rejected claims 12-17, 40, 42-48 and 50-55 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. (US 6,547,829) in view of Cho et al. ("Finding replicated web collections," hereinafter Cho et al);
- rejected claims 18-20, 37-39 and 56-58 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. (US 6,547,829) in view of Cho et al. and further in view of Rujan et al. (US 6,976,207); and
- rejected claim 49 under 37 U.S.C. §103(a) as being unpatentable over Meyerzon et al. (US 6,547,829) in view of Cho et al. and further in view of Lambert et al. (US Pub. No. 6,976,207).

After entry of this amendment, the pending claims are: claims 12-20, 37-40 and 42-58. No changes have been made to the claims.

Claim Rejections – 35 U.S.C. §103

Claims 12, 40, 50 and Associated Dependent Claims 13-17, 42-48, 49, and 51-55

The Examiner rejected claims 12-17, 40, 42-48 and 50-55 under 35 U.S.C. 103(a) as being unpatentable over Meyerzon and Cho. As acknowledged by the Examiner, Meyerzon does not teach or suggest the following limitations:

> "…indexing the representative document when the representative document is the newly crawled document; [and]
>
> "…repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective set of documents, such that **at least some of the** newly crawled documents are determined to be representative documents and are indexed."

See, Office Action, pg. 4. Instead, the Examiner relies on Cho to supply these limitations.

However, the Examiner's analysis of both Meyerzon and Cho fails to take into account the significance of the actual sequence of operations required by claim 12, and how

those operations result in the representative document for a set of documents changing from one document to another during the process. To facilitate this explanation, portions of claim 12 are reproduced here (with numbers inserted in front to facilitate reference to these elements in the discussion below):

> (2) "receiving a newly crawled document …;"
>
> (3) "… to identify a set of documents sharing the document identifier of the newly crawled document, and **ascertaining an original representative document** for the identified documents;"
>
> (4) "…updating the information …"
>
> (5) "…**determining a representative document for the newly crawled document and the identified set of documents**;"
>
> (6) "…indexing the representative document when the representative document is the newly crawled document; and"
>
> (7) "…repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective set of documents, such that **at least some of the newly crawled documents are determined to be representative documents and are indexed**."

In elements 2 and 3 of claim 12 reproduced above, after a newly crawled document is received, a set of documents that share the same document identifier as the newly crawled document are identified. Thus, claim 12 requires that the newly crawled document have the same content as the identified set of documents. Since this set of documents is identified in the claimed process, the documents in the identified set of documents are already known to the process (e.g., to a web crawler). In addition, "an original representative document for the identified documents" is ascertained in the third element reproduced above. This original representative document is "for the identified documents," and therefore the original representative document must be one of the identified documents, and is not the newly crawled document. This is important for the reasons explained next.

Next, in element 5 of claim 12, a representative document for the "newly crawled document and the identified set of documents" is determined. What does this mean? It means that the claimed process makes a new determination about which document is the representative document for those documents sharing the same document identifier as the newly crawled document.

Stopping right here, it can be shown that even these steps (2,3 and 5) of claim 12 define patentable subject matter over the combined teachings of Meyerzon and Cho. Both Meyerzon and Cho take an entire set of documents that has been determined to be identical or similar in content, and assign only one of those documents as the representative document. Neither Meyerzon nor Cho teach the operation of determining a (potentially new) representative document when a newly crawled document has the same document identifier (i.e., the same or similar content) as an identified set of documents that already has an original representative document. In fact, Cho teaches a process that avoids crawling and downloading documents that have the same or similar content to a representative document. In Cho, representative documents are identified after the completion of a crawl, and the accumulated information about a set of identical or similar documents is then used to avoid crawling any of the documents in the identified set other than the representative document during a new crawl. (*See* Cho, section 5.1, pages 21-22)

In Meyerzon, it is always the earliest crawled document of a set of identical (or similar) documents that is treated as the representative document. The other identical/similar documents are not indexed by Meyerzon. (*See* column 9, lines 18-50, and Figure 3. This point is further discussed in the next paragraph, below, with regard to the meaning of the term "indexing.") Thus, Meyerzon teaches away from and never performs the fifth element quoted above – that is, Meyerzon does not make a new determination of a representative document after crawling a document that is determined to be identical or similar to a set of identified, earlier crawled documents. (It is noted that the Examiner's citation to Meyerzon, column 9 lines 32-40, at the bottom of page 3 of the office action is not correct, because this portion of Meyerzon does not teach "determining a representative documents for the newly crawled document **and** the identified set of documents.")

Next, the sixth element reproduced above, says that the newly crawled document is indexed when it has been determined to be the new representative document, representative of all the documents that have the same document identifier (i.e., the same or similar content). What does this mean? It means that even though the crawler or search engine already knows about an original representative document, which has the same document identifier as the newly crawled document, the claimed process indexes the newly crawled document when it has been determined to be the representative document for this set of same/similar content documents. This is contrary to the teachings of Meyerzon. It is noted that the Examiner (top of page 4 of the office action) stated that Meyerzon does not teach this aspect of claim 12. Furthermore, looking at Figure 3, step S21 of Meyerzon, when Meyerzon

encounters a document having the same content as a previously known document (S21-Yes, Figure 3), Meyerzon simply stores the URL in a history table (S26) and does not index the document. As shown in Meyerzon at S21-No, a document is indexed (S25) ONLY WHEN the newly crawled document does not have the same content (same CID) as a previously known document. This point is also important because Meyerzon is using the term "indexing a document" in manner similar to claim 12, to mean that the content of the document is processed and added to an index (Index 400-1 in Meyerzon), and does not mean simply adding a URL to a history table.

With respect to Cho, the primary embodiments in Cho avoid crawling documents with the same content as a known representative document. However, as pointed out by the Examiner, section 5.2 of Cho discusses an embodiment in which duplicate documents are gathered. However, the Examiner's statement (page 4, 2nd paragraph) that Cho teaches indexing the representative document is not correct. A full reading of Cho reveals that Cho does not say that the duplicate documents are indexed. Rather, Cho teaches that when the representative document satisfies a search query, the search results include the representative document in a "rolled up" manner that includes a link to the representative document, and also a link (called the Replica link) for obtaining links to the duplicates (e.g., in case the representative document is not available from its host). *See* Cho § 5.2. Thus, like Meyerzon, Cho collects the URLs of documents that are believed to be the same or similar to a representative document, but that's it. No indexing.

Moving on to the seventh element reproduced above, the repeating operation of claim 12 requires that for at least one set of documents all having the same content, yet another representative document is determined upon receiving it from a web crawl, and then that document is indexed (i.e., its content is indexed). As explained above with respect to the fifth element reproduced above, changing the representative document from an original representative document to a newly crawled document is contrary to the teachings of both Meyerzon and Cho. In addition, as explained above with respect to the sixth element reproduced above, indexing the content of a document known to have the same content as a previously indexed document is also contrary to the teachings of both Meyerzon and Cho.

In summary, neither Meyerzon nor Cho teaches the fifth element of claim 12 reproduced above, because neither Meyerzon nor Cho ever identifies a newly crawled document as a representative document when other identical documents have been previously identified. In addition, neither Meyerzon nor Cho teaches the sixth element of claim 12 because neither Meyerzon nor Cho indexes the newly crawled document when the newly

crawled document is identified as the representative document. The seventh element of claim 12 is not taught by Meyerzon and Cho for the same reasons as the fifth and sixth elements.

It is noted that independent claims 40 and 50 have limitations corresponding to each of the limitations of claim 12 discussed above. For at least the reasons explained above, claims 12, 40, 50 and their dependent claims 13-17, 42-49, and 51-55 are patentable over the combined teachings of Meyerzon and Cho.

With respect to claim 49, it is noted that Lambert is cited only with respect to the use of temporary redirect pages, and does not teach the elements of claim 12 that are not taught by Meyerzon and Cho. Therefore claim 49 is patentable over the combined teachings of Meyerzon, Cho and Lambert for at least the same reasons as claim 40, from which claim 49 depends.

Claims 18, 37, 56 and Associated Dependent Claims 19-20, 38-39, 57-58

The Examiner rejected claims 18-20, 37-39 and 56-58 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon in view of Cho and in further view of Rujan. Rujan is cited solely for the proposition that it teaches removing an oldest table in a set of tables. Rujan does not teach anything about (A) identifying duplicate documents in a web crawl, (B) identifying a sequence of representative documents for a set of documents having the same content, and (C) determining which document or documents in a set of same/similar content documents to index.

It is noted claims 18, 37 and 56 all include limitations corresponding to each of the limitations of claim 12 discussed above. While these claims also include one or more additional limitations, the additional limitations do not need to be addressed here because all of these claims include at least three elements not taught by the combined teachings of Meyerzon, Cho and Rujan – i.e., the elements of claim 12 discussed in detail above. Therefore, for at least the reasons explained above, claims 18-20, 37-39 and 56-58 are patentable over the combined teachings of Meyerzon, Cho and Rujan.

With respect to claim 49, it is noted that Lambert is cited only with respect to the use of temporary redirect pages, and does not teach the elements of claim 12 that are not taught by Meyerzon and Cho. Therefore claim 49 is patentable over the combined teachings of Meyerzon, Cho and Lambert for at least the same reasons as claim 40, from which claim 49 depends.

CONCLUSION

In light of the amendments to the claims, and the arguments presented above, Applicants respectfully request that the Examiner reconsider this application with a view towards allowance. The Examiner is encouraged to call the undersigned attorney at (650) 843-4000 should any issues remain unresolved.

Respectfully submitted,

Date: October 22, 2007 / Gary S. Williams / 31,066

Gary S. Williams
MORGAN, LEWIS & BOCKIUS LLP
2 Palo Alto Square
3000 El Camino Real, Suite 700
Palo Alto, California 94306
(650) 843-4000